Amendment to the Claims

This listing of claims will replace all prior versions, and listings, of claims in the application:

Claim 1: (presently amended) A method comprising:

developing a language model from a tuning set of information;

segmenting at least a subset of a received textual corpus by clustering every N-items of the received corpus into a training unit, wherein resultant training units are separated by gaps;

calculating the similarity within a sequence of training chunks on either side of each of the gaps;

A8

selecting segment boundaries that maximize intra-segment similarity and inter-segment disparity; ~~and~~

calculating a perplexity value for each segment; and

refining the language model with one or more segments of the received corpus based, at least in part, on the calculated perplexity value for the one or more segments.

Claim 2: (original) A method according to claim 1, wherein the tuning set of information is application specific.

Claim 3: (original) A method according to claim 1, wherein the tuning set of information is comprised of one or more application-specific documents.

Claim 4: (original) A method according to claim 1, wherein the tuning set of information is a highly accurate set of textual information linguistically relevant to, but not taken from, the received textual corpus.

Claim 5: (original) A method according to claim 1, further comprising a training set comprised of at least the subset of the received textual corpus.

Claim 6: (original) A method according to claim 5, further comprising:

ranking the segments of the training set based, at least in part, on the calculated perplexity value for each segment.

Claim 7 (canceled)

Claim 8: (presently amended) A method according to claim 1 ~~7~~, wherein the resultant segment defines a training chunk.

Claim 9: (presently amended) A method according to claim 1 ~~7~~, wherein N is an empirically derived value based, at least in part, on the size of the received corpus.

Claim 10: (presently amended) A method according to claim 1 ~~7~~, wherein the calculation of the similarity within a sequence of training units defines a cohesion score.

Claim 11: (original) A method according to claim 10, wherein intra-segment similarity is measured by the cohesion score.

Claim 12: (presently amended) A method according to claim 10 ~~7~~, wherein inter-segment disparity is approximated from the cohesion score.

Claim 13: (original) A method according to claim 1 ~~7~~, wherein the calculation of inter-segment disparity defines a depth score.

A8

Claim 14: (original) A method according to claim 1, wherein the perplexity value is a measure of the predictive power of a certain language model to a segment of the received corpus.

Claim 15: (original) A method according to claim 1, further comprising:

ranking the segments of at least the subset of the received corpus based, at least in part, on the calculated perplexity value of each segment; and

updating the tuning set of information with one or more of the segments from at least the subset of the received corpus.

Claim 16: (original) A method according to claim 15, wherein one or more of the segments with the lowest perplexity value from at least the subset of the received corpus are added to the tuning set.

Claim 17: (original) A method according to claim 1, further comprising:

utilizing the refined language model in an application to predict a likelihood of another corpus.

Claim 18: (original) A storage medium comprising a plurality of executable instructions including at least a subset of which, when executed, implement a method according to claim 1.

Claim 19: (original) A system comprising:

a storage medium having stored therein a plurality of executable instructions; and

an execution unit, coupled to the storage medium, to execute at least a subset of the plurality of executable instructions to implement a method according to claim 1.

Claim 20: (presently amended) A storage medium comprising a plurality of executable instructions which, when executed, implement a language modeling agent to:

develop a language model from a tuning set of information;~~, to~~

segment at least a subset of a received textual corpus and calculate a perplexity value for each segment, wherein:

the language modeling agent segments the received corpus by clustering every N items of the received corpus into a training unit; and

the training units are separated by gaps;

calculate the similarity within a sequence of training units on either side of each of the gaps;

select segment boundaries that improve intra-segment similarity and inter-segment disparity; and ~~to~~

refine the language model with one or more segments of the received corpus based, at least in part, on the calculated perplexity value for the one or more segments.

Claim 21: (original) A storage medium according to claim 20, wherein the language modeling agent utilizes a tuning set of information relevant to that of the received corpus.

Claim 22: (original) A storage medium according to claim 20, wherein the language modeling agent ranks the segments of the training set based, at least in part, on a measure of similarity between two or more segments.

Claim 23 A storage medium according to claim 22, wherein the similarity measure is calculated for adjacent segments.

Claim 24 (canceled)

Claim 25: (original) A storage medium according to claim 20, further comprising instructions to implement an application which selectively invokes the language modeling agent to predict a likelihood of another corpus.

Claim 26: (original) A storage medium according to claim 25, wherein the application is one or more of a spelling and/or grammar checker, a word-processor, a speech recognition application, a language translation application, and the like.

Claim 27: (presently amended) A system comprising:

a storage medium drive, to removablye receive a storage medium according to claim 20; and

an execution unit, coupled to the storage medium drive, to execute at least a subset of the plurality of instructions and implement the language modeling agent.

Claim 28: (presently amended) A modeling agent comprising:

a controller, to receive invocation requests to develop a language model from a corpus; and

a data structure generator, responsive to the controller, to:

develop a language model from a tuning set of information;~~,~~

segment at least a subset of a ~~the~~ received corpus, wherein:

the segments of the received corpus are a clustering of every N items of the received corpus into a training unit; and

the training units are separated by gaps;

calculate the similarity within a sequence of training units on either side of each of the gaps;

select segment boundaries that improve intra-segment similarity and inter-segment disparity;

calculate a perplexity value for each segment;~~,~~ and

refine the language mode with one or more segments of the received corpus based, at least in part, on the calculated perplexity value.

Claim 29: (original) A modeling agent according to claim 28, wherein the tuning set is dynamically selected as relevant to the received corpus.

Claim 30: (original) A modeling agent according to claim 28, the data structure generator comprising:

a dynamic lexicon generation function, to develop an initial lexicon from the tuning set, and to update the lexicon with select segments from the received corpus.

Claim 31: (original) A modeling agent according to claim 28, the data structure generator comprising:

a frequency analysis function, to determine a frequency of occurrence of segments within the received corpus.

Claim 32: (original) A modeling agent according to claim 28, the data structure generator comprising:

a dynamic segmentation function, to iteratively segment the received corpus to improve a predictive performance attribute of the modeling agent.

AS

Claim 33: (original) A modeling agent according to claim 32, wherein the dynamic segmentation function iteratively re-segments the received corpus until the language model reaches an acceptable threshold.

Claim 34: (original) A modeling agent according to claim 32, the data structure generator further comprising:

a frequency analysis function, to determine a frequency of occurrence of segments within the received corpus.

Claim 35: (original) A modeling agent according to claim 34, wherein the data structure generator selectively removes segments from the data structure that do not meet a minimum frequency threshold, and dynamically re-segments the received corpus to improve predictive capability while reducing the size of the data structure.